



Universiteit
Leiden
The Netherlands

Automated 2-dimensional measurement of vestibular schwannoma: validity and accuracy of an artificial intelligence algorithm

Neve, O.M.; Romeijn, S.R.; Chen, Y.J.; Nagtegaal, L.; Grootjans, W.; Jansen, J.C.; ... ; Hensen, E.F.

Citation

Neve, O. M., Romeijn, S. R., Chen, Y. J., Nagtegaal, L., Grootjans, W., Jansen, J. C., ... Hensen, E. F. (2023). Automated 2-dimensional measurement of vestibular schwannoma: validity and accuracy of an artificial intelligence algorithm. *Otolaryngology - Head And Neck Surgery*, 169(6), 1582-1589. doi:10.1002/ohn.470

Version: Publisher's Version

License: [Creative Commons CC BY-NC 4.0 license](https://creativecommons.org/licenses/by-nc/4.0/)

Downloaded from: <https://hdl.handle.net/1887/3713780>

Note: To cite this publication please use the final published version (if applicable).

Automated 2-Dimensional Measurement of Vestibular Schwannoma: Validity and Accuracy of an Artificial Intelligence Algorithm

Olaf M. Neve, MD¹ , Stephan R. Romeijn, MSc² ,
Yunjie Chen, MSc³ , Larissa Nagtegaal, MSc^{1,2} ,
Willem Grootjans, PhD² , Jeroen C. Jansen, MD, PhD¹ ,
Marius Staring, PhD³ , Berit M. Verbist, MD, PhD² , and
Erik F. Hensen, MD, PhD¹ 

Otolaryngology–
Head and Neck Surgery
2023, Vol. 169(6) 1582–1589
© 2023 The Authors.
Otolaryngology–Head and Neck
Surgery published by Wiley
Periodicals LLC on behalf of
American Academy of
Otolaryngology–Head and Neck
Surgery Foundation.
DOI: 10.1002/ohn.470
<http://otojournal.org>
WILEY

Abstract

Objective. Validation of automated 2-dimensional (2D) diameter measurements of vestibular schwannomas on magnetic resonance imaging (MRI).

Study Design. Retrospective validation study using 2 data sets containing MRIs of vestibular schwannoma patients.

Setting. University Hospital in The Netherlands.

Methods. Two data sets were used, 1 containing 1 scan per patient ($n = 134$) and the other containing at least 3 consecutive MRIs of 51 patients, all with contrast-enhanced T1 or high-resolution T2 sequences. 2D measurements of the maximal extrameatal diameters in the axial plane were automatically derived from a 3D-convolutional neural network compared to manual measurements by 2 human observers. Intra- and interobserver variabilities were calculated using the intraclass correlation coefficient (ICC), agreement on tumor progression using Cohen's kappa.

Results. The human intra- and interobserver variability showed a high correlation (ICC: 0.98–0.99) and limits of agreement of 1.7 to 2.1 mm. Comparing the automated to human measurements resulted in ICC of 0.98 (95% confidence interval [CI]: 0.974; 0.987) and 0.97 (95% CI: 0.968; 0.984), with limits of agreement of 2.2 and 2.1 mm for diameters parallel and perpendicular to the posterior side of the temporal bone, respectively. There was satisfactory agreement on tumor progression between automated measurements and human observers (Cohen's $\kappa = 0.77$), better than the agreement between the human observers (Cohen's $\kappa = 0.74$).

Conclusion. Automated 2D diameter measurements and growth detection of vestibular schwannomas are at least as accurate as human 2D measurements. In clinical practice, measurements of the maximal extrameatal tumor (2D) diameters of vestibular schwannomas provide important

complementary information to total tumor volume (3D) measurements. Combining both in an automated measurement algorithm facilitates clinical adoption.

Keywords

artificial intelligence, automated measurement, vestibular schwannoma, volume

Received March 20, 2023; accepted July 11, 2023.

Vestibular schwannomas are benign intracranial tumors arising from the eighth cranial nerve. Patients typically present with audiovestibular symptoms such as hearing loss, balance problems, or tinnitus. Other symptoms include headache, facial paresis, or numbness.^{1–3} A small majority of vestibular schwannomas are nonprogressive, justifying active surveillance, with regular magnetic resonance imaging (MRI) as the preferred management strategy.⁴ However, some tumors are progressive, which ultimately can lead to brain stem compression or intracranial hypertension. To prevent these potentially life-threatening conditions, progressive tumors are usually treated with either radiotherapy or surgery.

¹Department of Otorhinolaryngology–Head and Neck Surgery, Leiden University Medical Center, Leiden, The Netherlands

²Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands

³Department of Radiology, Division of Image Processing, Leiden University Medical Center, Leiden, The Netherlands

Corresponding Author:

Olaf M. Neve, MD, Department of Otorhinolaryngology–Head and Neck Surgery, Leiden University Medical Center, Otorhinolaryngology H5-P, P.O. Box 9600, 2300 RC Leiden, The Netherlands.
Email: o.m.neve@lumc.nl

The accurate assessment of tumor progression is essential in clinical decision-making. Currently, tumor progression is determined based on the manual diameter measurements of subsequent MRIs.⁵ However, these measurements have considerable errors, with reported intra- and interobserver variabilities ranging between 10% and 40%.⁶⁻⁸ Compared to diameter measurements, volume measurements are considered to be more reliable for the detection of growth, however, these measurements are time-consuming.^{6,8,9} For that reason, volume measurements have not widely been adopted in clinical practice yet, neither manual nor semiautomated volume measurement algorithms.⁷

To overcome this problem, several fully automated volume measurement algorithms have been developed.¹⁰⁻¹³ These algorithms use deep learning techniques to determine tumor volume and show excellent performance compared to human volume measurements. The wider implementation of these algorithms has been hampered by the fact that they have been trained on single-center data, using single-vendor magnetic resonance (MR) scanners with limited variation in scan protocol. Therefore, the performance of these algorithms in different clinical settings is less reliable and requires additional external validation. We have recently developed an algorithm for the automated measurement of vestibular schwannomas that is based on multivendor, multicenter MR data, that has been validated externally and is applicable to different MR sequences.¹³

In current clinical practice, treatment decisions as well as consensus-based classifications such as those proposed by Koos et al¹⁴ and Kanzaki et al⁵ are not based on tumor volume but on extrameatal tumor diameters. Treatment decisions and tumor classifications focus on the extrameatal tumor parts rather than the whole tumor volume because the extrameatal extension is the closest proxy measurement to the anatomical relation and impact of the tumor to critical adjacent structures such as the brain stem.⁵ So, whereas volume change is superior in detecting tumor progression, extrameatal diameters provide essential additional information on the direction of tumor extension and progression. In 2018, a survey study showed that 91% of the members of the North American Skull Base Society would observe a small tumor (<15 mm cerebellar pontine angle [CPA]) until growth was detected.¹⁵ Since then, several papers have been published arguing for observation in small but progressive tumors (CPA < 15 mm), and a size threshold for active treatment was introduced, based on extrameatal tumor diameters, emphasizing the complementary value of tumor diameters to tumor volume measurements.^{16,17}

Therefore, this study aimed to validate an algorithm to measure extrameatal tumor diameters as an addition to a previously reported automated volume measurement algorithm.¹³ Combining automated 2-dimensional (2D) and tumor volume (3D) measurements in 1 algorithm would result in a robust tool suited to support treatment decisions in current clinical practice.

Methods

This retrospective study was performed in a University Hospital in The Netherlands, an expert center for vestibular schwannoma. The protocol has been reviewed by the Medical Research Ethics Committee Leiden Den Haag Delft (G19.115), which granted an exemption for informed consent.

Measurement Algorithm

This study aimed to extend the existing in-house developed automated volume measurement model with automated 2D measurements, that is, the maximal extrameatal tumor diameters in the axial plane. To do so, the automated 2D measurements were compared with repeated human measurements of 2 observers (O.M.N. and S.R.R.). The intra- and interobserver variability were analyzed. Second, the mean diameter of the 2 observers was used as ground truth to evaluate the automated measurements. All diameters were measured according to the consensus guidelines as proposed by Kanzaki et al,⁵ that is, the largest extrameatal diameter parallel to the petrous bone was measured first, followed by the largest extrameatal diameter perpendicular to the line drawn to acquire the first diameter (ie, perpendicular to the medial surface of the petrous bone).

The automated volume measurement algorithm, based on a convolutional neural network (CNN), was previously developed and validated by our research group¹³ using the nnU-net framework.¹⁸ For vestibular schwannomas, we used a 3D U-Net with 5 encoder and decoder layers, detailed in a previous publication by Neve et al.¹³ The model was trained and validated on scans from 37 different centers and was able to delineate tumors on contrast-enhanced T1 and on high-resolution (hr) T2.¹³ Furthermore, the performance was externally validated on the publicly available data set by Shapey and colleagues.^{13,19} In addition, the model was able to differentiate between the intra- and extrameatal tumor parts.

For the automated 2D measurements, the border between intra- and extrameatal tumor segmentations was used to select the plane parallel to the petrous bone, and orthogonal to the axial plane to mimic the clinical procedure. Using this plane, the largest parallel diameter was chosen from all axial slices in the segmentation. Consecutively, the largest diameter perpendicular to the parallel plane was derived from the same slice.

Design

Three analyses were performed. First, the intra- and interobserver variability of human 2D measurements was evaluated. Second, the accuracy of the automated 2D measurement was evaluated by comparing them to the human 2D diameters. Third, the capability to detect

tumor progression on consecutive scans based on automated 2D diameters was evaluated.

Study Population

Two different data sets were used in this study. The first was the data set used for the development of the automated segmentations from the study by Neve et al (development data set). This development data set contained 134 patients with 1 contrast-enhanced T1-weighted MRI. Of all MRIs the diameters were measured by 2 human observers (O.M.N. and S.R.R.) and in a subset of 50 patients both observers measured the diameters twice to assess the intraobserver variability.

Second, we randomly selected a data of 51 patients from vestibular schwannoma patients at our center, that had not been part of the first data set. These 51 patients had at least 3 consecutive MRIs without intercurrent active treatment (surgery or radiotherapy). This data set (the longitudinal data set) was used to assess tumor progression. Both observers (O.M.N. and S.R.R.) measured the diameters of all MRIs. In challenging cases, the observers consulted a senior head and neck radiologist (B.M.V.) with 22 years of experience to discuss the right plane and measurement. This consultation was performed in 6% of the MRIs. When contrast-enhanced T1 was not acquired, the measurement was performed on hrT2. Using both T1 and hrT2 mimics, the clinical setting in which either 1 or both sequences are used in follow-up.

For the evaluation of the intra- and interobserver variability of the human 2D measurements and the accuracy of the automated 2D measurements, both the development and longitudinal data sets were merged. Tumor progression analysis was performed on the longitudinal data set, as this contained multiple consecutive scans per patient.

Statistical Analysis

All analyses were performed in R version 4.1.1 using R-studio 1.4.1717 (Rstudio; PBC). The intra- and interobserver variability of human 2D measurements were evaluated by calculating the interclass correlation coefficient (ICC) and plotting Bland-Altman plots, containing the difference in measurement on the Y-axis and the mean of the measurements on the X-axis.²⁰ Bland-Altman limits of agreement were calculated by the mean difference between the measurements ± 1.96 times the standard deviation of the difference between measurements. CNN diameters were compared to the mean of the 2 human diameters to reduce the impact of human interobserver variability. Automated diameter outliers which exceeded the limits of agreement were analyzed by a senior head and neck radiologist (B.M.V.) and are discussed in the Discussion section.

Longitudinal tumor progression was based on a cutoff value of ≥ 2 mm difference between 2 consecutive scans. The mean of the 2 human measurements was used as

ground truth. CNN diameter progression performance was evaluated using sensitivity, specificity, and accuracy. In addition, Cohen κ was calculated. These results were compared to the agreement on tumor progression between the 2 human observers. The correlation of the maximal diameter in the axial plane (parallel or perpendicular) with the maximal diameter of the entire 3D extrameatal component was evaluated using the ICC.

Results

Patient characteristics of both data sets are shown in **Table 1** and technical characteristics in **Table 2**. In the longitudinal data set, 9 out of 153 scans could not be extracted from the picture archiving and communication system due to technical incompatibilities. The tumor size and cystic component distributions differ between the data sets. Patients in the first data set, used for the development of the automated volume CNN, were selected to have a large variety of tumor sizes. In contrast, the longitudinal data set was a random sample of all patients treated at our center. These selection methods might explain the difference in patient age since patients with larger tumors tend to be younger than patients with smaller tumors. Examples of the automated diameters are shown in **Figure 1**.

Intra- and Interobserver Variability

Interobserver differences of the human 2D measurements are shown in **Figures 2A** and **B**. The ICCs of the parallel and perpendicular measurements were both 0.984 (95% confidence interval [CI]: 0.976; 0.989); however, the limits of agreement were 1.7 mm and 1.9 mm, respectively.

Intraobserver differences provided similar ICCs for parallel (0.995, 95% CI: 0.992; 0.997) and perpendicular (0.989, 95% CI: 0.981; 0.993) measurements and the limits of agreement were 1.9 mm and 2.1 mm, respectively (shown in **Figures 2C** and **D**).

Table 1. Patient Characteristics

	Development data set	Longitudinal data set
N	134	51
MRI scans per patient	1	3
Age, y (SD)	53.5 (12.0)	61 (10.4)
Sex male	64 (48%)	28 (55%)
Cystic component	63 (47%)	7 (14%)
Tumor size		
Intrameatal	28 (21%)	20 (39%)
Small (0–10 mm)	19 (14%)	18 (35%)
Medium (11–20 mm)	26 (19%)	11 (22%)
Moderately large (21–30 mm)	24 (18%)	1 (2%)
Large (31–40 mm)	24 (18%)	1 (2%)
Giant (>40 mm)	13 (10%)	0

Abbreviation: MRI, magnetic resonance imaging.

Table 2. Technical Characteristics

	Development data set	Longitudinal data set	
	Contrast-enhanced T1-weighted MRI	Contrast-enhanced T1-weighted MRI	T2-weighted MRI
Number of scans	134	116	28
In-plane resolution	0.35 × 0.35 (0.27 × 0.27–1.0 × 1.0)	0.5 × 0.5 (0.27 × 0.27–1.13 × 1.13)	0.35 × 0.35 (0.20 × 0.20–0.55 × 0.55)
In-plane matrix	400 × 400 (256 × 208–560 × 560)	352 × 352 (256 × 192–640 × 520)	512 × 512 (256 × 256–1024 × 1024)
TE, ms	9 (2.38–20)	8.9 (2.38–22)	176.141 (1.968–263)
TR, ms	602.10 (8.76–2200)	450 (6.84–1900)	1200 (5.42–5110)
Section thickness	1.0 (0.9–5.0)	2 (0.6–6.0)	1 (0.5–3)

Abbreviations: MRI, magnetic resonance imaging; TE, time to echo; TR, repetition time.

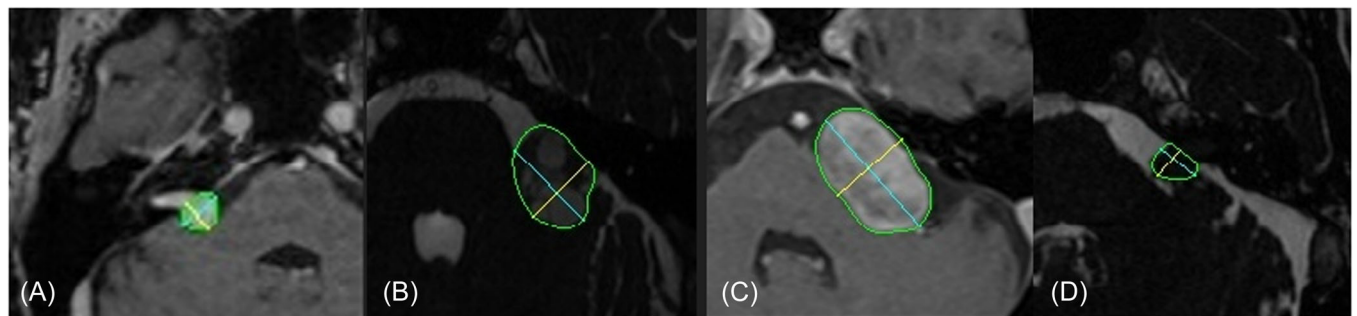


Figure 1. Automated diameter measurements on contrast-enhanced T1 (A, C) and hrT2 (B, D) MRI. Automated tumor segmentations (green line), largest extrameatal diameters parallel (blue line), and perpendicular (yellow line) to the petrous bone. hr, high resolution; MRI, magnetic resonance imaging.

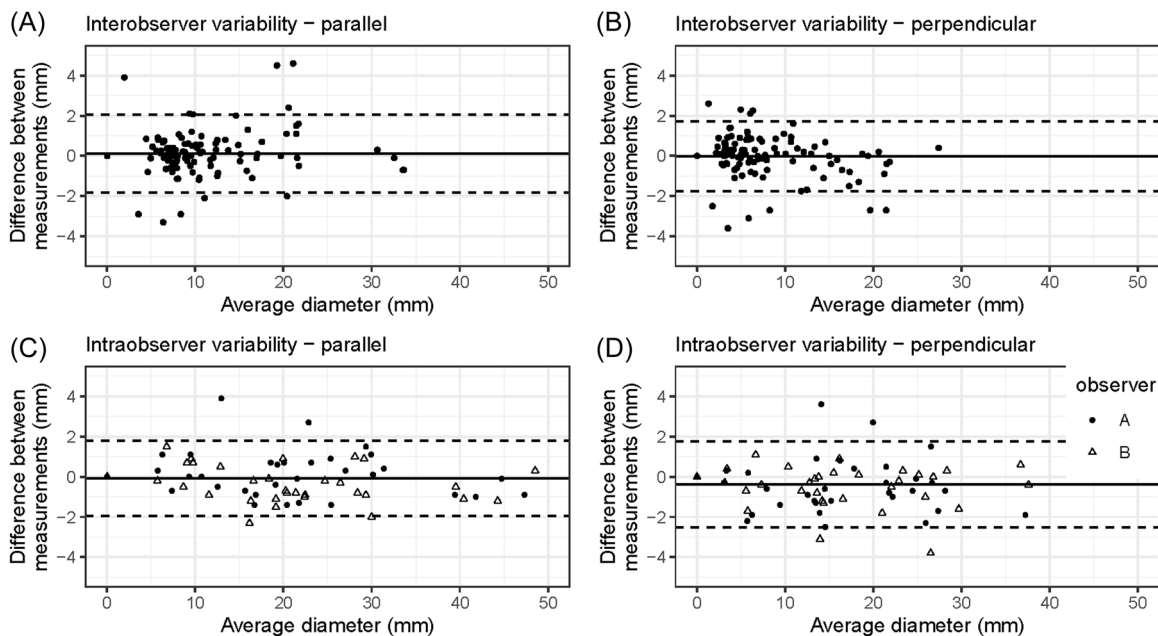


Figure 2. Bland-Altman plots of intra- and interobserver variability of human-derived diameter measurements (A–D). Limits of agreement (dotted line). The mean difference between measurements (black line).

Automated 2D measurement

The correlation between human and CNN diameters was excellent, with ICCs of 0.98 (95% CI: 0.974; 0.987) and 0.97 (95% CI: 0.968; 0.984) for the parallel and perpendicular diameters, respectively. As is shown

in **Figure 3**, the model diameters were, on average, slightly larger than the human diameters, resulting in a mean difference between human and CNN of 0.7 mm for parallel and 0.8 mm for perpendicular measurements. The limits of agreement were 2.2 mm for the

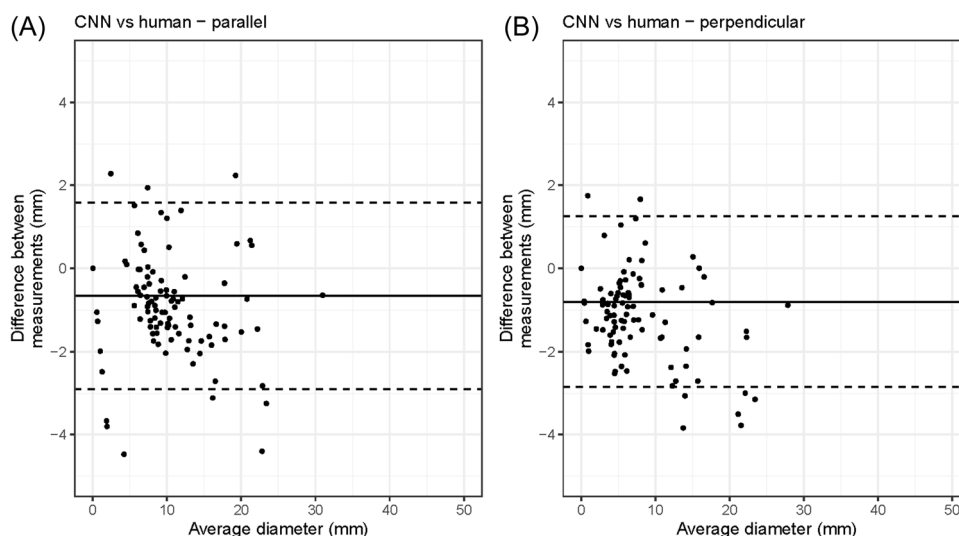


Figure 3. Bland-Altman plots of convolutional neural network (CNN) derived versus mean human-derived diameters (A, B). Limits of agreement (dotted line). The mean difference between measurements (black line).

parallel diameter and 2.1 for the perpendicular diameter.

Next, as the model is not confined to measurements in the axial plane, we evaluated the correlation of the maximal diameter in the axial plane (parallel or perpendicular) with the maximal diameter of the entire 3D extrameatal component. We found an excellent ICC of 0.974 (95% CI: 0.970; 0.984) between the largest diameter in the axial plane and the largest diameter in the entire 3D extrameatal component, as shown in **Figure 4**.

Tumor Progression

Table 3 shows the evaluation of agreement on the diameter progression of the CNN compared to the human measurements and agreement on the diameter progression of the 2 human observers. The agreement on tumor progression between the CNN and the mean of the 2 human observers resulted in a Cohen's κ of 0.77, indicating substantial agreement. Cohen's κ of the agreement between the 2 human observers was 0.74. Also, the sensitivity, specificity, and accuracy of the CNN compared to the mean of the 2 human observers were comparable to these values when comparing the 2 human observers.

Discussion

To our knowledge, this is the first study to propose an automated vestibular schwannoma 2D measurement algorithm using artificial intelligence techniques. The current study shows an intra- and interobserver measurement error of 1.7 to 2.1 mm in the 2D diameter measurement of vestibular schwannomas. The automated measurements were comparable to human measurements. The automated algorithm was able to detect tumor progression on consecutive MRI using either contrast-enhanced T1 or hrT2 sequences.

On average, the automated measurements were 0.7 to 0.8 mm larger than the human measurements. This difference may in part be caused by the fact humans decide by eyeballing what would be the maximal line to measure the diameter, while the automated method really maximizes this mathematically based on contrast differences. In addition, automated segmentations use contrast differences and maximize the segmentation on pixel level by including the contour lines of the tumor. Indeed, further analysis of outliers revealed that automatic measurements included the entire thickness of the segmentation contour line. Another explanation for the outliers was the difference between the algorithm and human observers in separating the intra- and extrameatal tumor parts. When a larger proportion of tumors is considered extrameatal, this affects the extrameatal diameters. The segmentation algorithm is trained on human segmentations of the whole tumor and the intra- and extrameatal tumor parts. The algorithm is not trained to detect specific anatomical structures such as the edge of the petrous bone, to determine the difference between intra- and extrameatal tumor parts. However, the use of other anatomical structures is incorporated indirectly since the human observers who annotated the training set did make use of the surrounding anatomical structures to determine the difference between the intra- and extrameatal tumor parts.

Both the intra- and interobserver variability of diameter measurements in vestibular schwannomas in the current study (respectively 0.98 and 0.99) are similar to previously reported ICCs. van de Langenberg et al⁸ and MacKeith et al⁷ have reported an ICC of 0.95 for interobserver agreement on diameter. Tolisano et al have reported a similar ICC of 0.98 and 0.99 for interobserver agreement on contrast-enhanced T1 and hrT2 sequences. The intraobserver variability has previously been described by MacKeith et al⁷ and Coelho et al²¹ ranging

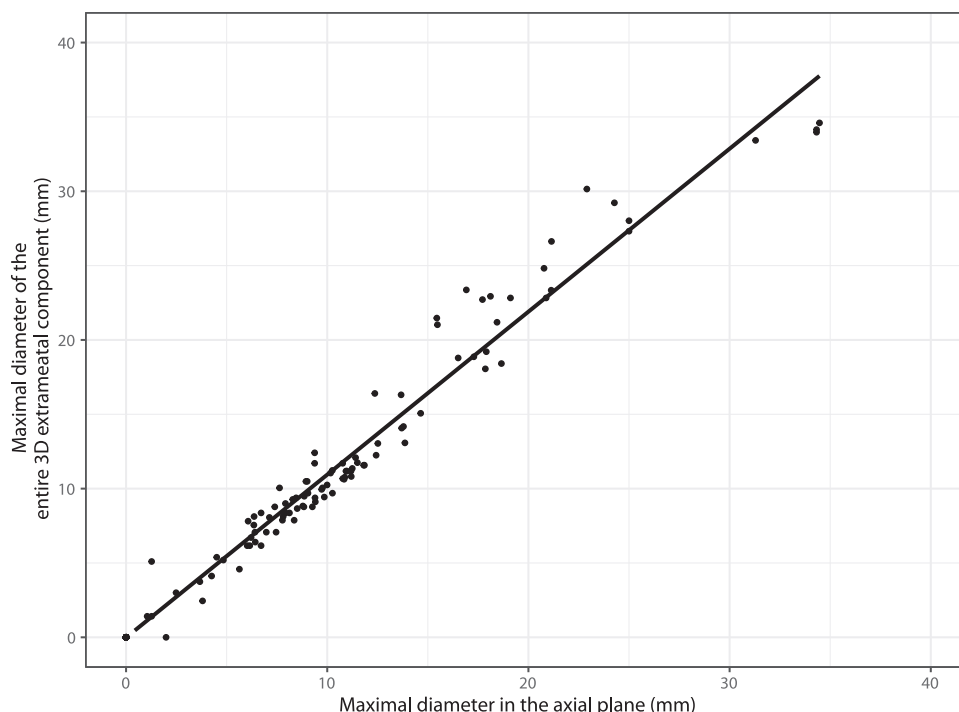


Figure 4. Correlation between the maximal extrameatal diameter in the axial plane with the maximal diameter of the entire 3-dimensional (3D) extrameatal component.

Table 3. Diameter Progression

	CNN vs human (≥ 2 mm = growth)	Observer 1 vs observer 2
Sensitivity	0.79	0.82
Specificity	0.90	0.90
Accuracy	0.86	0.88
Cohen κ	0.77	0.74

Abbreviation: CNN, convolutional neural network.

from 0.92 to 0.98. The ICC of the automated measurements compared to the mean of the 2 human measurements is similar at 0.98 and 0.97, indicating that the automated measurement is acceptable for use in clinical practice.

The study by Hougaard et al²² also used Bland-Altman limits of agreement for 2D measurement. They have reported limits of agreement for interobserver variability of 2.8 mm for parallel and 2.2 mm for perpendicular diameters. The intraobserver limits of agreement were smaller (2.6 and 1.9 mm). In the current study, the differences between interobserver (1.7 and 1.9 mm) and intraobserver (1.9 and 2.1 mm) limits of agreement were smaller and the interobserver limits were even lower compared to Hougaard et al. Considering the amount of variability in human diameter measurement, the performance of the automated diameter measurements (2.2 and 2.1 mm) is within the limits of human measurements.

The agreement on tumor progression based on diameter measurements on consecutive MRIs has been

analyzed by Tolisano et al using Cohen's κ . They reported a Cohen's κ of 0.56 and 0.61 for contrast-enhanced T1 and hrT2 sequences, respectively.²³ These agreement measures are slightly lower compared to Cohen's κ (0.74) found in the current study when the agreement between 2 human observers was compared. Automated diameter measurements (0.77) even outperformed this, showing the capabilities of the CNN to detect tumor growth.

This study has some limitations. As this analysis was performed on retrospective data, reliability needs to be validated using prospective data before use in clinical practice. In addition, the data set contained a small number of cystic tumors. These tumors are more challenging to delineate and could be prone to less accurate automated measurements. However, this is also true for manual measurements. Automated recognition of these cystic tumors could be a valuable improvement to the model as this could be used to alert radiologists to manually check the measurement of these tumors, thereby facilitating the clinical adoption of the tool. Furthermore, the data set also contained intrameatal tumors. Although this reflects clinical practice, the inclusion of intrameatal tumors was suboptimal for the validation of the automated extrameatal diameter measurements. Furthermore, the plane of the parallel extrameatal diameters was based on the border between intra- and extrameatal tumor parts. As a consequence, the algorithm was unable to measure the diameters of completely extrameatal tumors. In contrast, completely intrameatal tumors were detected and categorized with an extrameatal diameter of 0 mm.

Tumor diameter measurements show wide intra- and interobserver variability. Tumor volume measurements are widely accepted to more reliably detect tumor progression.⁸ However, volumetric measurements hold limited information about the direction of tumor extension. Furthermore, current consensus classification systems, such as those proposed by Kanzaki et al and Koos et al, are based on (extrameatal) diameter measurements. As the direction of the volumetric tumor progression is essential information in clinical decision-making, extrameatal diameters provide important information complementary to tumor volume (change). By including both measures in a reliable automated system that is able to deal with both contrast-enhanced T1 and hrT2 weighted MR imaging, we aim to provide a robust algorithm to support clinical decision-making in vestibular schwannoma patients.

The current algorithm is able to measure tumor diameters and volumes efficiently and consistently, which can be of added value in clinical practice compared to the currently used manual measurement limited to 2D diameters. Automated, consistent measurement of both diameters and volumes in consecutive scans could improve the accuracy of tumor growth detection as well as provide therapy-relevant information while saving time and costs. It could therefore be a useful and efficient tool for multicenter vestibular schwannoma research and care; however, future research is needed to evaluate the impact of incorporating automated tumor measurements and progression detection on clinical practice.

Conclusion

The accuracy of automated 2D measurements is comparable to manual 2D diameter measurements. Adding 2D diameters to tumor 3D volume measurements in 1 automated model provides a robust algorithm that can assist in clinical decision-making in vestibular schwannoma patients. The algorithm proposed in this study is able to deal with both contrast-enhanced T1 and hrT2 weighted MR imaging of different MR scanner types and protocols, enabling its use in a multicenter setting.

Author Contributions

Olaf M. Neve, study concept and design, data collection and analysis, interpretation, and implications of results, drafting manuscript, and final approval of the manuscript; **Stephan R. Romeijn**, study concept and design, data collection and analysis, interpretation, and implications of results, and final approval of the manuscript; **Yunjie Chen**, study concept and design, data collection and analysis, interpretation and implications of results, and final approval of the manuscript; **Larissa Nagtegaal**, study concept and design, interpretation, and implications of results, and final approval of the manuscript; **Willem Grootjans**, study concept and design, interpretation, and implications of results, and final approval of the manuscript; **Jeroen C. Jansen**, study concept and design, interpretation, and implications of results, and final approval of the manuscript;

Marius Staring, study concept and design, interpretation, and implications of results, and final approval of the manuscript; **Berit M. Verbist**, study concept and design, interpretation, and implications of results, and final approval of the manuscript; **Erik F. Hensen**, study concept and design, interpretation, and implications of results, drafting the manuscript, and final approval of the manuscript.

Disclosures

Competing interests: The authors declare that they have no competing interests.

Funding source: This research was funded by a strategic fund of the Leiden University Medical Center and 1 of the authors (Y.C.) was funded by a China Scholarship Council Grant (No. 202008130140).

Data Availability Statement

Data generated or analyzed during the study are available from the corresponding author by reasonable request.

ORCID iD

Olaf M. Neve  <http://orcid.org/0000-0002-5104-8448>
 Stephan R. Romeijn  <http://orcid.org/0000-0002-4634-447X>
 Yunjie Chen  <http://orcid.org/0000-0001-9478-6953>
 Larissa Nagtegaal  <http://orcid.org/0000-0002-2618-0228>
 Willem Grootjans  <http://orcid.org/0000-0003-4851-7167>
 Jeroen C. Jansen  <http://orcid.org/0000-0002-3955-0152>
 Marius Staring  <http://orcid.org/0000-0003-2885-5812>
 Berit M. Verbist  <http://orcid.org/0000-0002-1010-2583>
 Erik F. Hensen  <http://orcid.org/0000-0002-4393-7421>

References

1. Arthurs BJ, Fairbanks RK, Demakas JJ, et al. A review of treatment modalities for vestibular schwannoma. *Neurosurg Rev.* 2011;34:265-279. doi:10.1007/s10143-011-0307-8
2. Management of sporadic vestibular schwannoma. *Otolaryngol Clin North Am.* 2015;48:407-422.
3. Matthies C, Samii M. Management of 1000 vestibular schwannomas (acoustic neuromas): clinical presentation. *Neurosurgery.* 1997;40:1-9.
4. Møller MN, Hansen S, Miyazaki H, Stangerup SE, Caye-Thomasen P. Active treatment is not indicated in the majority of patients diagnosed with a vestibular schwannoma: a review on the natural history of hearing and tumor growth. *Curr Otorhinolaryngol Rep.* 2014;2:242-247. doi:10.1007/s40136-014-0064-7
5. Kanzaki J, Tos M, Sanna M, Moffat DA. New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma. *Otol Neurotol.* 2003;24(4):642-649. doi:10.1097/00129492-200307000-00019
6. Varughese JK, Wentzel-Larsen T, Vassbotn F, Moen G, Lund-Johansen M. Analysis of vestibular schwannoma size in multiple dimensions: a comparative cohort study of different measurement techniques. *Clin Otolaryngol.* 2010;35(2):97-103. doi:10.1111/j.1749-4486.2010.02099.x
7. Mackeith S, Das T, Graves M, et al. A comparison of semi-automated volumetric vs linear measurement of small

- vestibular schwannomas. *Eur Arch Otolaryngol*. 2018;275(4):867-874. doi:10.1007/s00405-018-4865-z
8. van de Langenberg R, de Bondt BJ, Nelemans PJ, Baumert BG, Stokroos RJ. Follow-up assessment of vestibular schwannomas: volume quantification versus two-dimensional measurements. *Neuroradiology*. 2009;51(8):517-524. doi:10.1007/s00234-009-0529-4
9. Cross JJ, Baguley DM, Antoun NM, Moffat DA, Prevost AT. Reproducibility of volume measurements of vestibular schwannomas—a preliminary study. *Clin Otolaryngol*. 2006;31(2):123-129. doi:10.1111/j.1749-4486.2006.01161.x
10. Shapey J, Wang G, Dorent R, et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. *J Neurosurg*. 2021; 134:171-179. doi:10.3171/2019.9.JNS191949
11. Lee C, Lee W-K, Wu C-C, et al. Applying artificial intelligence to longitudinal imaging analysis of vestibular schwannoma following radiosurgery. *Sci Rep*. 2021;11(1):3106. doi:10.1038/s41598-021-82665-8
12. George-Jones NA, Wang K, Wang J, Hunter JB. Automated detection of vestibular schwannoma growth using a two-dimensional U-Net convolutional neural network. *Laryngoscope*. 2021;131(2):131. doi:10.1002/lary.28695
13. Neve OM, Chen Y, Tao Q, et al. Fully automated 3D vestibular schwannoma segmentation with and without gadolinium-based contrast material: a multicenter, multi-vendor study. *Radiol Artif Intell*. 2022;4(4):e210300. doi:10.1148/ryai.210300
14. Koos WT, Day JD, Matula C, Levy DI. Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas. *J Neurosurg*. 1998;88(3):506-512. doi:10.3171/jns.1998.88.3.0506
15. Carlson M, Van Gompel J, Wiet R, et al. A cross-sectional survey of the north american skull base society: current practice patterns of vestibular schwannoma evaluation and management in North America. *J Neurol Surg B Skull Base*. 2018;79(3):289-296. doi:10.1055/s-0037-1607319
16. Macielak RJ, Wallerius KP, Lawlor SK, et al. Defining clinically significant tumor size in vestibular schwannoma to inform timing of microsurgery during wait-and-scan management: moving beyond minimum detectable growth. *J Neurosurg*. Published online October 15, 2021. doi:10.3171/2021.4.jns21465
17. Marinelli JP, Lohse CM, Carlson ML. Introducing an evidence-based approach to wait-and-scan management of sporadic vestibular schwannoma. *Otolaryngol Clin North Am*. 2023;56(3):445-457. doi:10.1016/j.otc.2023.02.006
18. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
19. Shapey J, Kujawa A, Dorent R, et al. *Data From: Segmentation of Vestibular Schwannoma From Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm* [Dataset]. The Cancer Imaging Archive; 2021. doi:10.7937/TCIA.9YTI-5Q73
20. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327(8476):307-310.
21. Coelho DH, Tang Y, Suddarth B, Mamdani M. MRI surveillance of vestibular schwannomas without contrast enhancement: clinical and economic evaluation. *Laryngoscope*. 2018;128(1):202-209. doi:10.1002/lary.26589
22. Hougaard D, Norgaard A, Pedersen T, Bibby BM, Ovesen T. Is a redefinition of the growth criteria of vestibular schwannomas needed. *Am J Otolaryngol*. 2014;35(2):192-197. doi:10.1016/j.amjoto.2013.08.002
23. Tolisano AM, Wick CC, Hunter JB. Comparing linear and volumetric vestibular schwannoma measurements between T1 and T2 magnetic resonance imaging sequences. *Otol Neurotol*. 2019;40:67. doi:10.1097/mao.0000000000002208